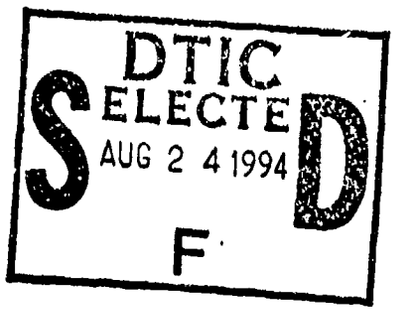
NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
NAVAL AIR STATION, PENSACOLA, FL 32508-5700

NAMRL-1363

AN EVALUATION OF PERFORMANCE-BASED TESTS DESIGNED TO IMPROVE NAVAL AVIATION SELECTION

D. J. Blower and D. L. Dolgin



Reviewed and approved 9 August 1991

J. A. Brady

J. A. BRADY, CAPT, MSC, USN
Commanding Officer

This research was sponsored by the Naval Medical Research and Development Command under work units 63706N M0096.001-7007 and 63706N M0096.001-7054.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Trade names of materials and/or products of commercial or nongovernment organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such commercial materials and/or products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1991	Interim Apr 86 - Dec 89

4. TITLE AND SUBTITLE
An Evaluation of Performance-based Tests Designed to Improve Naval Aviation Selection

5. FUNDING NUMBERS
63706N
M0096.002 M00960.01

6. AUTHOR(S)
David J. Blower
Daniel L. Dolgin

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Aerospace Medical Research Laboratory
Bldg. 1953, Naval Air Station
Pensacola, FL 32508-5700

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRL-1364

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center, Bldg. 1
Bethesda, MD 20889-5044

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report describes the evaluation of a portion of a new aircrew selection test battery recently developed at the Naval Aerospace Medical Research Laboratory. The results indicate that performance-based test measures can be used to predict flight training performance. Several test measures were reliably related to a pass/fail criterion. These results provide support for the prediction of whether or not a candidate will pass or fail training. The results of a hierarchical multiple regression revealed that scores from three tests, 1) Absolute Difference-Horizontal Tracking, 2) Complex Visual Information, and 3) Risk-taking Task, were generally equivalent in predicting success in primary flight training. Interactions of college major and accession source with derived scores of the three significant tests contributed significant amounts of variability when added to the model. We recommend that the valid tests from this study be implemented for operational use with the AQT/FAR. The use of hierarchical multiple regression with the tests will isolate those specific measures capable of accounting for added and unique variance, beyond that of the AQT/FAR and certain demographic variables, in the prediction of primary flight training course.

14. SUBJECT TERMS
aviation selection, attrition, prediction, flight training, performance-based, psychomotor tests, psychological tests

15. NUMBER OF PAGES
37

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

SUMMARY PAGE

THE PROBLEM

We have developed an automated performance-based test battery over the past several years in order to augment the Navy's paper-and-pencil selection tests for aviators. Whereas previous reports have discussed various individual tests, this report contains a detailed description of all the tests that comprise the current test battery. We had two basic reasons for conducting this evaluation, one scientific in nature and the other grounded in economic concerns. First, the entire battery had never been used in a linear regression equation model to predict success in primary flight training. Secondly, we needed to reduce the existing battery for the practical reasons of transferring the tests to an operational setting .

FINDINGS

This analysis revealed that derived scores from three tests, (1) Absolute Difference—Horizontal Tracking, (2) Complex Visual Information Processing (CVT), and (3) a Risk-Taking Task, were generally equivalent in predicting success in primary flight training. The derived scores from the Manikin, Baddeley, and Psychomotor/Dichotic Listening Task tests did not account for any significant variance. In addition, the linear regression models were *not* improved by adding the variables of other test sets when the model already included one significant test set. Interactions of college major and accession source with derived scores of the three significant test sets predicted significant amounts of variability when added to the model.

RECOMMENDATIONS

We recommend eliminating some tests from the existing battery for the purpose of transitioning the tests to an operational setting. While any of the three tests identified above could remain in the test battery, the analysis showed that the CVT test resulted in the best model for prediction when moderated by certain demographic variables. This model indicates that using the CVT in the selection process would not be indicated for Naval Academy graduates and those college graduates with engineering and math majors. As a final recommendation, we encourage the use of hierarchical multiple regression as

a statistical technique. The effects of obvious, easy-to-obtain predictor variables must be accounted for before assessing the effects of newly proposed tests that purport to improve the prediction of success in flight training. This is especially true when new tests are expensive to implement, and other predictor variables are already available, or are relatively inexpensive to obtain. From our experience with this data set we also recommend examining models that contain *interactions* of the easy-to-obtain variables with one or two promising tests. These interaction terms may be able to explain as much, or perhaps even more, variance as performance on additional proposed tests.

Acknowledgments

The authors gratefully acknowledge the technical assistance of Peter D. Collyer for hardware and software contributions to the test battery and for his persistence in obtaining the criterion measures. The diligent and conscientious efforts of Mr. Al Thomas, HM2 Solomon Eagles, Ms. Sylvia Starling, and Ms. Becky Dodson are also acknowledged in test administration. We would be remiss in not mentioning some of our predecessors who laid the groundwork for much that is reported here: Dr. Gerry Gibb, Dr. Lee Goodman, Mr. Ray Griffin, LCDR Dennis McBride, LCDR Tommy Morrison, and CDR Jerry Owens. We are much indebted to Prof. Harold Delaney of the University of New Mexico for his professional interest, advice, suggestions, and overall contribution to the work reported here.

The research reported in this paper was completed under the Naval Medical Research and Development Command work unit numbers 63706N M0096.001-7007 and 63706N M0096.001-7054.

The views expressed in this report are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. Trade names of materials and/or products of commercial or non-government organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such commercial materials and/or products. Volunteer subjects were recruited, evaluated and employed in accordance with the procedures specified in Department of Defense Directive 3216.2 and Secretary of the Navy Instruction 3900.39 series. These instructions are based upon voluntary informed consent and meet or exceed the provisions of prevailing national and international guidelines.

INTRODUCTION

In 1986, the Naval Aerospace Medical Research Laboratory (NAMRL) completed the development of an automated performance assessment test battery to be used with the Navy's primary screening instrument: the paper-and-pencil Academic Qualification Test/Flight Aptitude Rating (AQT/FAR). This battery was designed to assess cognitive abilities, higher-order processes, psychomotor skills, time-sharing ability, and personality traits that might predict success in naval aviation training.

The criteria used to select each test in the battery was based on previous studies [1]–[5]:

1. uniqueness of other tests in the battery
2. its potential for automation
3. minimal test administration requirements
4. relevance to aircrew
5. construct validity

Recently, the Air Force, in a validation study [6,7], has investigated the performance-based tests that comprise the Basic Attributes Battery (BAT). The BAT is intended for use with the Air Force Officer Qualification Test (AFOQT), a paper-and-pencil test similar to the Navy's AQT/FAR, for aircrew selection and classification purposes. The Air Force study, with 883 subjects, identified several performance-based tests that would improve aircrew selection if used with the AFOQT. These predictors included time sharing, spatial, cognitive, and personality measures as selected by a stepwise regression model.

This report describes the final validation for primary flight training of the Navy experimental aviation selection test battery developed at NAMRL. The test battery was "designed to complement the BAT to avoid duplication of effort" [4]. Within the past few years, several of the tests in the NAMRL selection battery have been evaluated and reported individually [8]–[14]. In this report we provide a first look at all the tests combined in one model, and focus on the unique contribution of each test.

METHODS

SUBJECTS

Student naval aviators, preselected for naval aviation flight training on the basis of their performance on the current Navy and Marine Corps aviation selection tests and medical examinations, participated in the study. The subjects were informed that a)

the investigation involved performing tasks in problem solving and perceptual and motor skills, b) that their test performance would not affect their continuation in the program nor be entered into their permanent service records and that c) results would be used solely for the purpose of developing an improved aviation selection program for the Navy. The student naval aviator candidates were 20-30 years old ($M = 23.29, SD = 1.51$) and averaged 24.95 previous civilian flight hours ($SD = 136.79$). The vast majority of the subjects in our overall data base were male. For example, out of 1,110 cases, only 21 were female.

Several classification variables were recorded for each subject. These included:

1. initial aviation selection test scores
2. previous civilian flight training
3. age
4. source of procurement
5. sex
6. college major

Previous civilian flight training was treated as a continuous variable using the self-reported number of total flight hours a subject had logged. Total flight hours included both solo and dual pilot training hours. Source of procurement included six distinct groups:

1. Aviation Officer Candidate School (college graduates entering directly into the military)
2. Naval Academy graduates
3. Naval Cadets (prior enlisted service or 60 college credits with numerous other criteria)
4. Marine Corps Officer Program
5. Naval Reserve Officer Training Corps
6. Other (direct procurement, Merchant Marine Academy, enlisted commissioning programs)

College major was classified into one of five general disciplines:

1. engineering and math
2. physical sciences (biology, geology, physics, etc.)
3. business
4. social sciences (psychology, sociology, history, etc.)
5. physical education

APPARATUS

All testing was conducted on Apple IIe microcomputers with Apple monochrome monitors (CRTs). Subjects used a numeric keypad to respond to discrete stimuli. All responses were recorded to millisecond accuracy. A Measurement Systems Incorporated (MSI-542) control stick was used for joystick and throttle control during the tracking tasks. Rudderpedal controls were measured using a variable resistor connected to a computer A/D channel. The joystick was mounted on the forward edge of the testing console at a centered position. The throttle was located on the left side of the testing booth. The rudderpedals were located so that the subjects could easily place both feet while sitting in the testing booth. Subjects operated the joystick with the right hand, the throttle with the left hand, and the rudderpedals with both feet.

PROCEDURE

All candidates were tested before entering flight training and after completing a 14-week basic military indoctrination program for AOC officers or a 6-week program for students already commissioned. All instructions were presented to the subjects on the CRT for each task individually. Test administrators intervened only to begin the computer program for each task and to answer questions posed by subjects. The test administration time of the battery ranged from 3.7 to 4.0 h. The order of the tasks and the stimuli within each test were identical for all subjects. Subjects received a 3-4 min rest period between tasks. All testing took place in an air-conditioned laboratory.

PREDICTOR TESTS

In today's military aviation environment, aircrew personnel are often required to monitor several tasks concurrently. It is not unusual for the pilot to perform communication and navigation functions, while at the same time be faced with the physical demands of controlling the aircraft. The NAMRL battery was designed to measure and in some cases simulate the skills required of an operational pilot. Appendix A provides a summary of the NAMRL selection test battery and lists each test, the time (in minutes) to take the test, and a brief description of the attributes measured.

For all of the tests described below (except for the AQT/FAR), each subject was asked to respond as quickly and accurately as possible. The number of correct and incorrect responses and their associated reaction times were recorded. Because of problems with the capacity of the online data storage system, only summary measures were obtained for the Baddeley Test of Grammatical Reasoning. The subject always indicated a "correct"

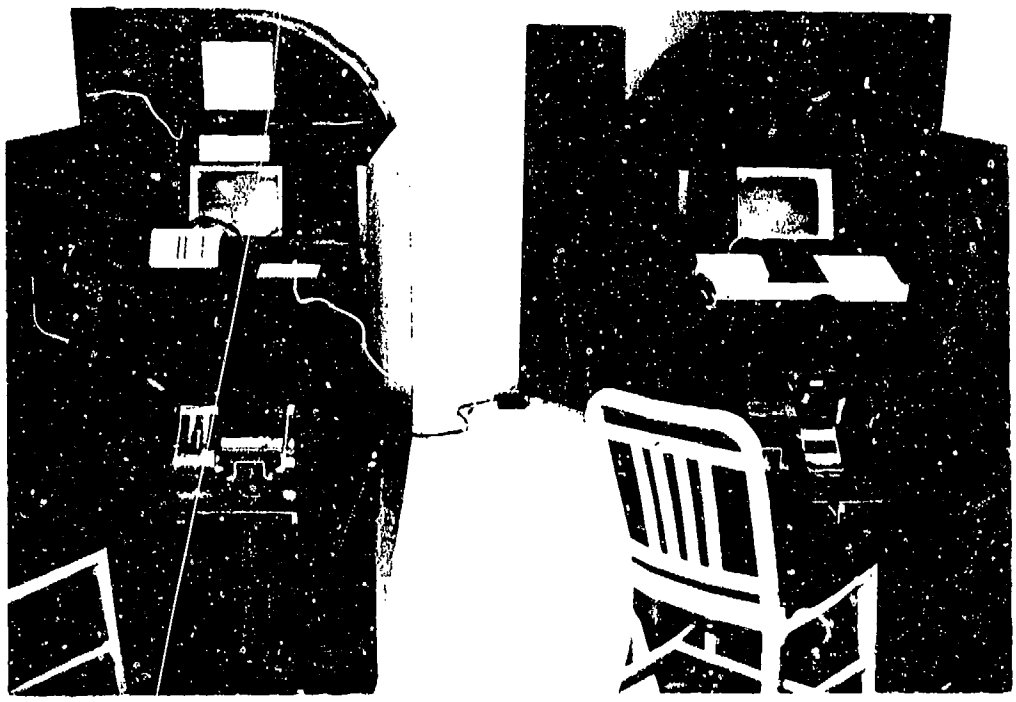


Figure 1: Two views of the automated performance-based test apparatus.

or "same" response by pressing the key under the right index finger for all tests. An "incorrect" or "different" response was indicated by pressing the key under the right second (middle) finger. The subjects received no performance feedback on any of the tasks, with the one exception of the risk-taking test. All predictor tests analyzed were self-paced, with the exception of time-limit impositions placed on the dichotic listening task (DLT) and tracking tasks. Rest periods were scheduled within and between tests to decrease any effects of mental or physical fatigue. Figure 1 displays two views of the test apparatus.

ACADEMIC QUALIFICATION TEST/FLIGHT APTITUDE RATING (AQT/FAR)

Presently, the U.S. Navy, Marine Corps, and Coast Guard use the AQT/FAR for pilot selection. The AQT/FAR is a paper-and-pencil test limited to the assessment of academic aptitude, interests, mechanical comprehension, and background variables. There is no performance-based measurement or evaluation included in the selection test battery. The battery has remained essentially the same since its inception 47 years ago.

The AQT/FAR is not part of NAMRL's performance-based test battery. The AQT/FAR

is administered to flight training applicants by Navy recruiters worldwide. A passing score on the AQT/FAR is required for entrance into flight training. All subjects participating in this study were preselected for flight training partially on the basis of their AQT/FAR scores. These scores were provided to NAMRL by the Naval Aerospace Medical Institute which is the custodian of the AQT/FAR data base.

AQT

The AQT assesses a candidate's level of intellectual or academic ability. The test consists of 105 items measuring abilities in vocabulary, practical judgment, reading comprehension, mathematics, and arithmetic reasoning. Another ability assessed is that of making detailed comparisons between long, complex character strings. Practical judgment is tested by presenting complicated situations in which each alternative requires the assessment of resulting gains and losses. This is a relatively standard form of aptitude test that resembles those encountered in high school and college settings (*e.g.*, the Scholastic Aptitude Test and the Graduate Record Examination).

MECHANICAL COMPREHENSION TEST (MCT)

The MCT contains 75 items to assess skills and abilities in the following areas: (a) spatial-mechanical relationships, (b) factual-mechanical information, and (c) systems analysis. The test is relatively demanding so an applicant must either know the mechanical principles or memorize the specific examples to do well.

SPATIAL APPERCEPTION TEST (SAT)

This 34-item test assesses the ability to abstract from a personal (first person) view to a detached (third person) view. A pilot's view of the forward horizon is given in one picture frame, followed by five frames depicting a ground view that best corresponds to the aircraft represented by the pilot's view. Four examples are given, and the underlying decision rules are presented.

BIOGRAPHICAL INVENTORY (BI)

The BI is a 160-item test that provides measures for positive and negative personal attributes for aviation officer selection. The questions cover personal history (e.g., membership in clubs) personal health (e.g., occasions of illness), attitudes (e.g., preferences in coursework), and aviation knowledge (e.g., propellant types for specific rocket engines). Unlike the other AQT/FAR subtests, the BI is untimed. The instructions to the applicant indicate that there are no right or wrong answers, and that they may be checked for accuracy in a follow-up interview.

FAR

The FAR is a composite of scores derived from the MCT, two-dimensional SAT, and BI. Stanine scores from 1 to 9 are derived for both the AQT and FAR. Typically, an AQT of 3 and a FAR of 5 are required for entry into naval flight training, although entry requirements vary among accession sources and reflect current Navy needs. The AQT/FAR requires 4h to complete.

PSYCHOMOTOR TESTS (PMT)

The PMT consists of stick, rudder pedal, and throttle control tasks. Subjects are required to maintain computer-generated cursors on designated center points of the CRT. The subjects control horizontal cursor movement using their right hand on the joystick and both feet on the rudder pedals. The throttle controls are gripped by the left hand to control vertical movement of the cursor. The computer automatically measures pixel distar rror) from actual cursor location to cursor target positions on the CRT. Cursor move was in the same direction for the throttle tasks but in the opposite direction for the k and rudder pedal tasks.

The PMT becomes progressively more difficult over a series of three sessions. The first session requires the subject to control a single cursor (keeping it in the center of the CRT) with the joystick only. The second session required subjects to maintain two cursors (one for stick and one for rudder pedal controls) on separate central CRT positions. The final and most complex session requires that the subject manipulate and maintain simultaneously the joystick, rudder pedal, and throttle cursor controls . The PMT requires 55 min testing time and assesses eye-hand-foot coordination.

DICHOTIC LISTENING TASK (DLT)

The DLT was constructed to study individual differences in the ability to focus attention [15]. In the DLT, subjects are instructed to selectively attend to one of two messages that are presented simultaneously, one to each ear. The DLT is a 20 min test that assesses divided attention by a series of letters and digits presented to each ear. "Left" and "right" commands direct the subject to attend to one ear while ignoring the other ear. The subject reports the sequence either verbally or electronically by keypad entry, depending on test sequence.

PMT/DLT COMBINATION

Certain studies [16,17] have demonstrated that time-sharing abilities correlate with flight performance although they are not tested by the DLT as a single instrument. To

study this factor and provide a measure of multiple-task performance, the PMT was combined with the DLT to assess shared resources. As mentioned above, the PMT and DLT are administered initially as single tasks and then combined to result in increasingly complex, simultaneously performed, multiple-task conditions. Testing time for the combined PMT/DLT is approximately 75 min.

MANIKIN TEST

Because spatial processing plays a critical role in aviator performance, we included the manikin test in the battery as a measure of visual-spatial performance. Previous research has demonstrated that the manikin test has been found to be significantly correlated with standardized paper-and-pencil tests of spatial abilities [18,19].

The manikin test consists of 48 drawings, 5cm by 3cm, of a sailor holding a square in one hand and a triangle in the other. The sailor is depicted either right side up or upside down, facing toward or away from the subject. The objective is to quickly determine which hand is holding the square. The test measures spatial orientation and reaction time. Each subject completes eight 2-min trials separated by a 20-s rest period. Test length is 16-min.

HORIZONTAL TRACKING

In this task, the subject is required to learn a one-dimensional compensatory tracking task. To perform the task, the subject must anticipate the movement of a cursor on the computer screen and manipulate the joystick to counterbalance the movement in order to keep the cursor centered on a fixed central point on the screen. For example, if the cursor is moving off center to the right the subject would move the joystick to the left in order to re-center the cursor. Specifically, the subject maintains a 0.6-cm square centered in a 9.75 by 1.25-cm rectangle by moving the joystick either left or right. The cursor is driven by a forcing function programmed into the computer. Each subject receives ten 2-min trials separated by a 30-s rest. The dependent measure is RMS error. Total testing time is 20-min. Figure 2 displays a photograph of the horizontal tracking task as it appears to the subject on the computer.

ABSOLUTE DIFFERENCE TASK

In this task, the subject is visually presented with a random number between 1 and 9 on the CRT, which is erased and then followed immediately with another number. This task is essentially a measure of short-term memory, memory search, and encoding. The subject is required to press the numeric key (keys 1 through 4), that represents the absolute difference between the number presently displayed on the CRT screen and the

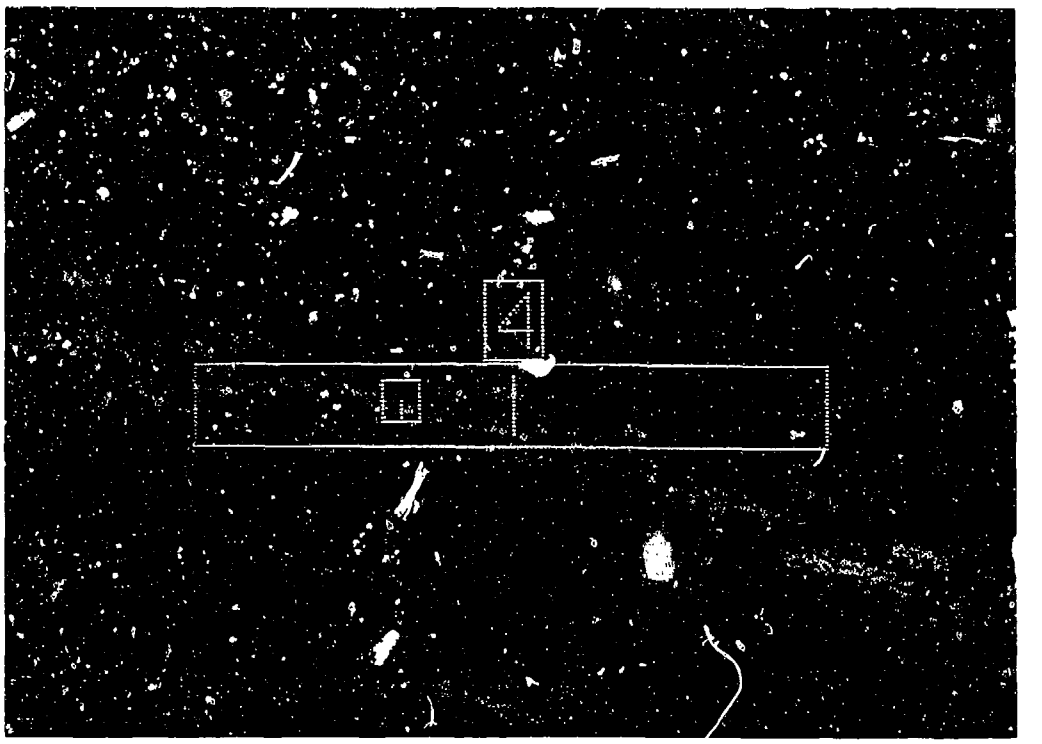


Figure 2: The combined one-dimensional tracking task and absolute difference task.

number shown on the screen on the last trial. When a response is entered, a new number appears on the CRT, whereby the subject calculates the absolute difference between the number previously presented and the number currently displayed. The subject is instructed to use only his right hand in responding with a numeric keypad. Speed and accuracy of response are emphasized. The task is subject-paced and consists of fifteen 2-min trials.

ABSOLUTE DIFFERENCE AND HORIZONTAL TRACKING

As part of the emphasis on time-sharing ability within the NAMRL battery, the horizontal tracking and absolute difference tasks are combined for a measure of dual-task performance. For this combination of tasks, the subject performs the horizontal tracking task and the absolute difference task simultaneously. The stimuli for the absolute difference task are centered above the tracking task and touch the top of the tracking task. The subject controls the tracking task with his right hand and, for the absolute difference task, presses a number on the keypad with his left hand. The subject is told that the two tasks are equally important. The subject receives five 2-min trials separated by rest periods of 30-60s. Testing time for the dual-task combination is 50 min.

COMPLEX VISUAL INFORMATION PROCESSING TASK (CVT)

The CVT is a test of visual/perceptual abilities. After an introduction and specific examples of what will be encountered on the test, a question is presented on the CRT regarding the position of different objects on a slide. The slide is projected only after the subject indicates (via the keypad) that the question is understood. The subject's task is to memorize detailed questions about individual slides that will be presented later in the test. A total of 120 slides are presented that contain different "symbols" that represent airplanes, aircraft carriers, and destroyers. For example, a triangle might represent an airplane; a rectangle a destroyer. In addition to the various symbols, each slide is divided into quadrants with vertical and horizontal axes.

The following is a typical CVT testing sequence. First, the subject is presented with a question on the CRT (e.g., how many aircraft carriers are heading north into the upper right quadrant). After the subject reads the question and presses the "enter" key on the computer keypad, the question disappears from the monitor, and the slide is presented depicting the various symbols. To accurately respond to the prompt, the subject must remember the question and indicate the correct response regarding the number of carriers heading north within the specified segment of the slide. In other words, the CRT presents the question asking how many carriers (triangle symbol) did you see on the slide? A slide is then presented that displays a diagram of symbols. The subject must recall and respond to symbols configuration and location. Subjects are told that reaction time and number of correct responses are crucial for accurate measurement on the task.

A total of three measures are derived from the CVT: 1) the time required to read the question, 2) the time to respond to the display, and 3) the number of correct responses. The CVT is the lengthiest test in the battery and requires 50 min to administer. A photograph of a subject performing on the CVT is shown in Fig. 3.

BADDELEY GRAMMATICAL REASONING

Baddeley's reasoning task [20] was included in the battery as a measure of logical reasoning. This task consisted of each subject describing the order of presentation of two letters, A and B, either correctly or incorrectly. Subjects are required to determine whether various simple sentences and their grammatical transformations correctly describe the relational order of two alphabetic letters. For example, "A follows B...BA." The subject determines if the sentence correctly describes the order of the two letters. The subject is then required to make a speeded true/false judgment by pressing the (1) key ("true") or the (2) key ("false"). The task is subject-paced and lasts approximately 10 min.

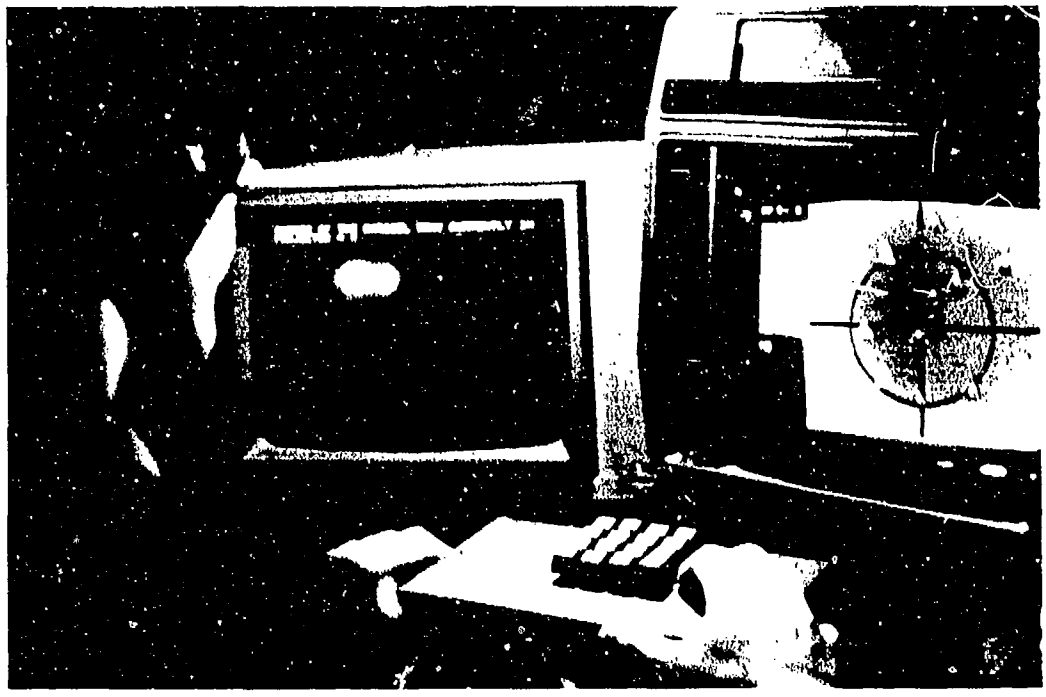


Figure 3: A photograph of a subject being tested on the CVT task.

RISK TEST

The risk test is based on a gambling scenario developed from a test described by Slovic [21]. An initial version of this instrument was subsequently revised at NAMRL [8]. The original test had 3 sessions of 10 trials each, but initial research indicated that the first session had the highest correlation with the performance criteria, so only the first session was retained for continued evaluation. Figure 4 provides a photograph of the risk test as it appears on the CRT.

For each of the 10 trials of the test session, the subject views a 2 by 5 matrix of squares with each cell of the matrix filled consecutively by the numbers 1 through 9 and 0 (to indicate 10). At the beginning of each trial, one square is randomly chosen by computer as a penalty square, and the remaining nine become reward squares. Any of these 10 squares can then be selected by pressing keys corresponding to the numbers in the squares. Because the location of the reward and penalty squares are randomly selected by the computer at the beginning of each trial, subjects have no prior knowledge as to where the reward and penalty squares are located. The subject can select any one of the squares, and if the selected square is a reward square, the subject receives a payoff. The subject retains these payoffs in a cumulative fashion with the total payoff amount for that trial indicated on the screen during the trial. If the subject selects the penalty

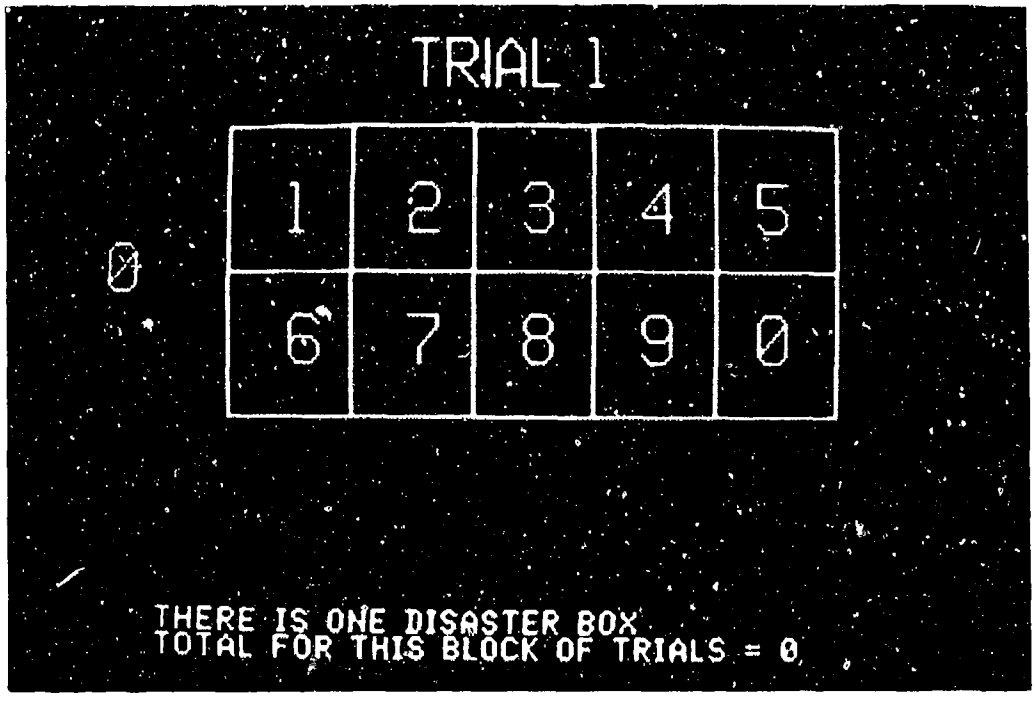


Figure 4: A photograph of the risk test as it appears on the computer screen.

square, however, the trial ends, the total payoff for that trial is lost, and the next trial begins. Subjects can stop at any time during a trial, retain the payoff accumulated, and go automatically to the next trial. The instructions given to the subjects are neutral in tone regarding the number of boxes to select and how quickly to respond.

The average number of squares selected per trial, as well as the average latency (for each trial), are calculated and retained for each session/subject pair. The test lasted approximately 15 min with no task learning time. Final scores for analysis were mean number of responses (NR) per trial corresponding to squares accumulated and response times (RT) for those NRs. Points are directly related to NRs, therefore only NRs are analyzed. Increased risk taking on this instrument is indicated by increases in number of responses made and/or decreases in response times.

RESULTS

The SPSS/PC+ statistical software package was used to analyze the data on a Zenith Z-248 (IBM PC AT compatible machine). Hierarchical multiple regression was performed using the procedure REGRESSION from SPSS/PC+.

Appendix B contains a description of the data base used in this analysis as well as summary statistics for the derived scores. The formulas detailing how the derived scores were obtained from the original raw variables are also presented in Appendix B.

The motivation for employing a hierarchical approach to regression is lucidly explained by Cohen and Cohen [22] and by Tabachnick and Fidell [23]. Essentially, in using hierarchical regression we attempted to "control for," "parcel out," or "adjust for" the influence of certain variables on the criterion before we ascertain the effects of the main variablesin question. In our study, we sought to statistically control for the effects of demographic variables like age and college major first, then adjust for intelligence by parcelling out the effects of AQT/FAR. This elimination was done before we looked at any possible influence of the performance-based test battery on success in primary flight training.

In essence, we subtracted the influence of the inexpensively obtained predictors before we judged the merits of an expensive set of predictor variables. To prove worthy of consideration, the expensive set of predictors had to add something above and beyond what is already accounted for by more easily gathered predictors.

The "inexpensive" set of predictors consisted of 11 demographic and intelligence variables while the "expensive" set of predictor variables consisted of 25 derived scores from the performance-based test battery for a total of 36 variables in all. Differing numbers of student naval aviators entered into the analysis depending on the particular set of test scores under consideration. The number of subjects for which we had scores on all the variables that entered the regression equation ranged from a low of 337 on the RISK test to a high of 1077 for the DEMO and AQT/FAR set of variables.

The criterion was a dichotomous variable indicating pass or failure in primary flight training. A pass was coded as "1" and a failure was coded as "0". Of the many ways to fail,we had 21 different attrition codes for academic, motivational, physical qualification, and other reasons for failure. For our purposes all different failure codes were lumped together.

Three special problems must be considered because the criterion variable is binary.

1. The error terms are not normally distributed.
2. The error variance is not a constant.
3. The output of a model may lie outside the interval zero to one.

The three solutions to these problems are:

1. Weighted least squares can be employed to estimate the parameters [24].
2. A nonlinear model using the logistic function can be attempted [25].
3. An ordinary unweighted least squares may be attempted ignoring the special problems listed above.

Unweighted least squares for a linear function was employed in this analysis because the intent was to conduct an initial screening of the tests, not to obtain a refined estimate

Table 1: A list of the predictor variables, their abbreviations, and their order of entry in the hierarchical multiple regression.

Entry Order	Test Set	Abbreviation
1	Demographic	DEMO
2	Academic Qualification Test & Flight Aptitude Rating	AQT/FAR
3	Complex Visual Information Processing	CVT
3	Risk	RISK
3	Absolute Difference & Horizontal Tracking	ADHT
3	Manikin and Baddeley	MB
3	Psychomotor/Dichotic Listening	PMT/DLT

of the parameters. The estimated regression coefficients will still be unbiased, but these estimates may have an inflated error variance. Also, because the probability of a pass in primary flight training is known from historical records to be about 90% (borne out in our data sample), the error variance is less affected than if we had equal numbers of pass and fail in the sample analyzed.

In hierarchical regression, the predictor variables enter the regression equation in a specified order. Table 1 shows the variables that entered into the equation and when they were entered. In this study, the predictor variables entered the regression equation according to the following "variance stealing" logic.

The set of five demographic variables was entered first. This set accounted for as much variance in the pass/fail criterion measure as possible. The set of six AQT/FAR variables was entered next. They accounted for as much variance as possible in the pass/fail criterion left over by the demographic set. Five separate regression equations were then computed. In each of these five equations, a set of scores from a performance-based test was added to the sets of demographic and AQT/FAR variables. In this manner, we could determine how much unique variance was accounted for by a particular test.

Table 2 shows the results of this analysis. The first column shows which test set is in the equation. The last five rows have the DEMO and AQT/FAR sets already in the equation. The second column shows the change in R^2 when that particular test set has been entered. This is the squared semipartial correlation coefficient (sr_i^2). The squared semipartial correlation coefficient represents the variance accounted for in the pass/fail criterion by the test in question *after* the variance that the demographic and intelligence

Table 2: Results of a hierarchical multiple regression to assess the relative contributions of various test sets.

Test Set	sr_i^2	Adjusted R^2	N	# Variables
DEMO	.01526*	.01066	1077	5
AQT/FAR	.01329*	.01852	1077	6
CVT	.03752*	.05064	557	3
ADHT	.03432*	.03289	499	8
RISK	.03369*	.05028	337	2
MB	.01814	.02479	544	5
PMT/DLT	.00987	.01037	641	7

$*p < .05$

sets *have in common* with the test score set have been removed. The third column shows the adjusted R^2 for the regression equation. The adjusted R^2 takes into account the differing sample sizes and number of parameters in the regression equations and places all the equations on an equal footing. The fourth column contains the number of subjects, and the fifth column contains the number of variables in the test set.

In using hierarchical multiple regression, we want to see how much each test can contribute when the demographic and intelligence variables have been held constant. Another way of stating this is that we want to assess differences in pass/fail performance that are due solely to differences in test score performance and not to differences in age, sex, accession source, college major, prior flight hours, or intelligence.

From Table 2 we see that the CVT, ADHT, and the RISK sets of predictor variables were the only test scores which accounted for significant additional variance after controlling for the demographic and intelligence variables. Each one of these sets accounted for about 3.5% of the variance in the pass/fail criterion.

Additional hierarchical multiple regressions were run in an attempt to discover if, after one set of test variables was part of the equation, another set would add additional significant variance as measured by the R^2 change. These results were all negative. All attempts to generate a regression equation using more than one test set to find a higher adjusted R^2 also failed.

Therefore, these results indicate that, after controlling for certain demographic and intelligence variables, each of three tests in the performance-based test battery (CVT, ADHT, and RISK) predict about 3.5% of the variability of the dichotomous success variable for primary flight training. The predictive ability of the regression equations

was *not* improved by adding any further test sets to the ones mentioned above.

In search of an even better regression model, we tested certain interactions. These interactions were entered as a fourth set of variables in the hierarchy after DEMO, AQT/FAR, and the set of variables for a test were entered. Table 3 shows the results of this analysis. The first column shows the particular interaction tested, the second column the change in R^2 resulting from the addition of the interaction set to the already existing equation, the third column shows the adjusted R^2 for the full model with interactions, and the final column shows the total number of degrees of freedom consumed by the parameters in the full model with interaction terms.

Significant changes in R^2 were caused by the following five interaction sets:

1. AGE by MB
2. ACCESSION by CVT
3. ACCESSION by RISK
4. ACCESSION by ADHT
5. EDUCATION by CVT

Table 3 reveals that the squared semi-partial correlation coefficient for the accession source by ADHT interaction term was significant, and that the squared semi-partial correlation coefficient for the college major by ADHT interaction term, while relatively large, was *not* significant. Eight derived scores from the original 54 variables make up the ADHT test set. This means that a large number of degrees of freedom are consumed for any regression model that includes the ADHT test set, especially when interaction terms are studied. The adjusted R^2 is also adversely affected by the large number of parameters that need to be estimated.

Therefore, a principal components analysis of the original 30 ADHT variables was conducted using procedure FACTOR from SPSS PC+ to derive a smaller set of ADHT variables. The first four principal components, accounting for 77.8% of the variance, were extracted from the 30 original ADHT variables. Factor scores were then computed for the first four principal components for all subjects who took the ADHT test.

A hierarchical multiple regression was then performed using these four derived scores instead of the previous eight to reduce the number of degrees of freedom due to regression sources in the model. This regression model included accession source by ADHT (principal components) and college major by ADHT (principal components) interaction terms. Table 4 shows the results of this analysis in the form of an ANOVA table.

Tables 5 and 6 show the ANOVA tables for the regressions run on the CVT and RISK test sets and their interaction terms. Multiple R, multiple R^2, the adjusted R^2, and the change in R^2 due to the addition of the interaction terms (R^2 Δ) are presented. The F values for all three models are significant showing that the multiple Rs are significantly different from zero. The R^2 Δs indicate that adding the interaction terms to the three models results in a significant amount of variability that is predicted. We reemphasize

Table 3: Results of the hierarchical multiple regression with selected interactions included in the equation.

Test Set	sr_i^2	Adjusted R^2	# Variables
AGExCVT	.00402	.04951	17
AGExRISK	.01596	.06107	15
AGExADHT	.01135	.03461	24
AGExMB	.03414*	.05096	21
AGExPMTDLT	.00531	.00786	23
SEXxCVT	.00810	.05371	17
SEXxRISK	.00042	.04480	15
SEXxADHT	.00041	.02313	24
SEXxMB	.00015	.01748	20
SEXxPMTDLT	.00556	.00812	23
ACCESSIONxCVT	.02762*	.06975	21
ACCESSIONxRISK	.02985*	.06782	18
ACCESSIONxADHT	.05795*	.05057	36
ACCESSIONxMB	.02653	.03280	27
ACCESSIONxPMTDLT	.01573	.00501	33
EDUCATIONxCVT	.02301*	.06121	21
EDUCATIONxRISK	.00717	.04112	18
EDUCATIONxADHT	.04005	.04089	36
EDUCATIONxMB	.02279	.02780	27
EDUCATIONxPMTDLT	.02186	.01002	33

$*p < .05$

Table 4: ANOVA Table showing final hierarchical multiple regression model for ADHT principal component scores. The model includes college major by ADHT and accession source by ADHT interaction terms.

Source	*df*	*Sum of Squares*	*Mean Square*	*F*
Regression	33	5.78	.175	1.89*
Residual	465	43.15	.093	
$R = .34379$		$R^2 = .11819$		
$adj.R^2 = .05561$		$R^2\ \Delta = .05789$**		

$*p < .0025$
$**p < .02$

that these models contain all DEMO variables, all AQT/FAR variables, all particular test set variables, and, in addition, the relevant interaction terms. The adjusted R^2s are used to rank order the models by compensating for the varying sample sizes and varying number of parameters estimated in the different models.

Because we coded the criterion variable of pass or fail in primary flight training, we can determine the "probability of success" for a group as defined by the ratio of those who pass over the number in the entire group. In Table 7, we show the probability of success in a representative case for three important factors highlighted by our regression results. These factors are accession source, college major, and number of correct answers on the CVT test.

The accession source factor has three levels:

1. Aviation Officer Candidate (AOC) and Officer Candidate School (OCS) graduates.
2. U.S. Naval Academy (USNA) graduates.
3. Others: ROTC, Naval Cadet (NAVCAD), U.S. Marine Corps (USMC), and those who did not fall into the existing categories.

Three levels of the college major factor are:

1. Engineering/Mathematics majors
2. General Science majors
3. Business, Humanities, Social Science, and Physical Education majors.

The CVT score factor was arbitrarily divided into three levels by selecting cutoff scores for the number of correct answers on the CVT test. These three levels are labelled low, medium, and high.

Table 5: ANOVA Table showing final hierarchical multiple regression model for CVT scores. The model includes college major by CVT and accession source by CVT interaction terms.

Source	*df*	*Sum of Squares*	*Mean Square*	*F*
Regression	28	7.41	.265	2.79*
Residual	528	50.01	.095	
$R = .35924$		$R^2 = .12906$		
$adj.R^2 = .08287$		$R^2\ \Delta = .05326$**		

$*p < .0001$
$**p < .01$

Table 6: ANOVA Table showing final hierarchical multiple regression model for RISK scores. The model includes an accession source by RISK interaction term.

Source	*df*	*Sum of Squares*	*Mean Square*	*F*
Regression	18	4.24	.236	2.36*
Residual	318	31.77	.099	
$R = .34316$		$R^2 = .11776$		
$adj.R^2 = .06782$		$R^2\ \Delta = .02985$**		

$*p < .002$
$**p < .05$

Table 7: Probability of success for accession source, college major, and number of correct answers on CVT test.

Factor	Probability of success	Number in group
Accession Source		
AOC/OCS	.85	280
Naval Academy	.92	114
Other	.91	164
College Major		
Eng/Math	.94	241
Gen. Science	.87	76
Other	.83	241
CVT Score		
Low	.80	123
Medium	.87	237
High	.94	198

Because our best regression model includes interaction terms of CVT with accession source and CVT with college major we show the probability of success for the breakdown of CVT scores and accession source in Table 8, and CVT scores and college major in Table 9. The number of subjects comprising each of these groups is shown in parentheses below the respective probability of success. To more vividly illustrate the clear effect of these interactions, the numbers in Tables 8 and 9 are plotted in Fig. 5.

The main contributor to the interaction of accession source with CVT is the behavior of the USNA graduates. Their probability of success is relatively flat as a function of CVT scores, whereas the other two groups exhibit roughly parallel curves showing increasing probability of success as CVT scores go from "low" to "high."

The same situation occurs in the interaction of CVT with college major at the top of the graph shown in Fig. 5. The General Science majors and Other majors are parallel, but the Engineering and Mathematics majors show a much more gentle rise in probability of success as a function of increasing CVT score.

DISCUSSION

We employed hierarchical multiple regression to assess the relative merits of the individual tests comprising our test battery to predict success in primary flight training.

Table 8: Probabili y of success broken down by CVT score and college major.

College Major	CVT Score		
	Low	Medium	High
Engineering/Mathematics	.91 (34)	.93 (105)	.95 (102)
General Science	.74 (23)	.89 (27)	.96 (26)
Business/Humanities Social Science/PE	.77 (66)	.81 (105)	.93 (70)

Table 9: Probability of success broken down by CVT score and accession source.

Accession Source	CVT Score		
	Low	Medium	High
AOC/OCS	.76 (71)	.83 (125)	.96 (84)
Naval Academy	.93 (15)	.92 (50)	.92 (49)
ROTC/USMC/NAVCAD/Other	.84 (37)	.92 (62)	.94 (65)

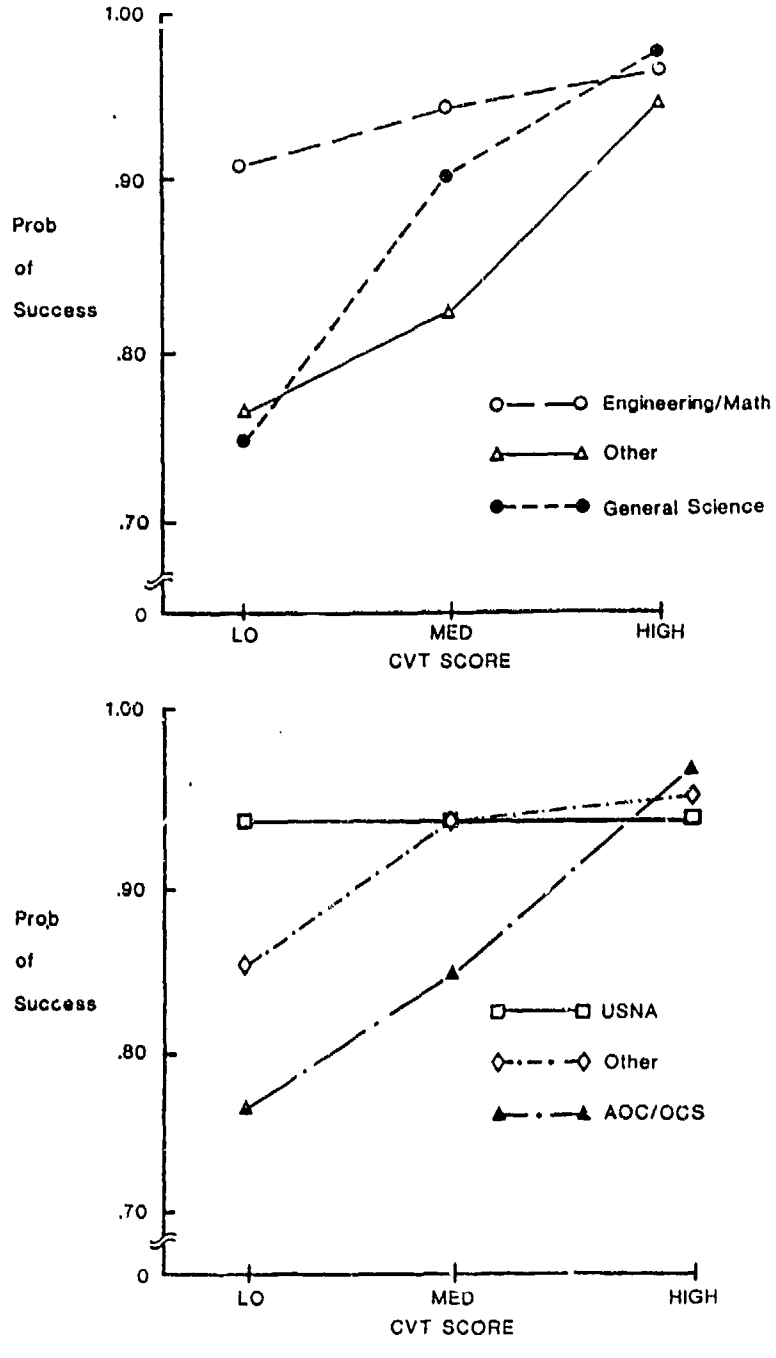


Figure 5: Graphs of the interactions of accession source and college major with correct answers on the CVT test.

Hierarchical multiple regression equalizes the tests by statistically controlling for certain demographic and intelligence variables.

This analysis revealed that derived scores from three tests, ADHT, CVT, and RISK were generally equivalent in predicting success. The derived scores from the Manikin—Baddeley (MB) and Psychomotor/Dichotic Listening Task (PMT/DLT) did not account for any significant variance after controlling for the demographic and intelligence variables. In addition, the linear regression models were *not* improved by adding the variables of other test sets when the model already included one significant test set.

On the other hand, interactions of college major and accession source with the variables of the three significant test sets were shown to contribute significant amounts of variability when added to the model. These results thus appear to indicate differential validity of these selection tests.

Using the adjusted R^2 as a standard, the "best" linear regression model discovered so far is one that includes 7 demographic variables (accession source and college major were each transformed to 2 dummy variables to account for the 3 levels), 6 AQT/FAR variables, 3 CVT scores, and 12 interaction terms. This model accounts for about 8.3% of the variance of the dichotomous criterion measure.

These results are in general agreement with similar studies conducted by the Air Force in their selection research program. Carretta [26], in a cross-validation study, reported correlations in the range of .20 – .24 for predicting success in USAF undergraduate pilot training. These correlations, when squared, represent an estimate of 4–5% variability explained by the Air Force tests.

If one assumes that using an adjusted R^2 as an estimate of explained variability is generally equivalent to the cross-validation squared correlations, then these two independent studies support the notion that perhaps 5–10% of the variability in primary flight training outcome can be explained using these selection tests. In our case, to achieve this level of explained variance, we need to include interactions of certain demographic variables with the performance on the test.

We underscore the point that these models were developed from a restricted population, that is, student naval aviators who had already passed medical and aptitude tests, and who had successfully completed several weeks of general military indoctrination. If these tests were administered to a more general population to predict success in primary flight training, we could expect even better performance from these selection models.

We believe that these regression models represent a good *initial* characterization of the data. We do *not* claim that we have necessarily found the very best models. For example, dropping some of the variables may result in a more parsimonious model, and there may very well be other, less obvious, interactions lurking in the data. We would be more comfortable with regression estimates stemming from a weighted least squares and/or a

logistic transformation analysis. These refinements, however, must await a subsequent treatment.

CONCLUSIONS

This study supports downsizing the existing battery so it is more manageable in an operational setting. Any one of the three tests identified could be a candidate for transition, and only one of the three would be sufficient. The tests and the administration times are as follows:

1. RISK—10 min
2. ADHT—75 min
3. CVT—50 min

A linear regression equation has been developed for each one of these tests and the equations could be utilized to select naval aviators with increased cost savings over the present selection methods. There seems to be reasonable evidence for an estimate in the range of 5–10% reduction in error variability.

The unearthing of significant test score interactions with procurement source and college major is another major finding of this study. These interactions point to a case of differential validity if any of the three tests are used for selection purposes. It is desirable, therefore, to develop separate regression equations for selection purposes depending, for example, on whether the pilot candidate is a Naval Academy graduate or is an AOC graduate. Interestingly, student pilots from the Naval Academy who have graduated from preflight training have almost half the attrition rate (7.4% vs. 12.3%) of AOC preflight graduates [27]. This means that our selection models must be viewed in an even better light because it is, in fact, the AOC group for which we can better predict success. That is, our selection models do a better job for exactly that group of students who have a higher attrition rate during flight training.

In the past, the US Air Force has excluded graduates from the USAF Academy from testing on the AFOQT (the Air Force's equivalent to the Navy's present selection test). Although this decision to exempt Air Force Academy graduates may initially appear arbitrary, based on our current findings, the Air Force policy seems to rest on a solid rationale.

References

[1] Melton, A. W., *Apparatus Tests*, Army Air Forces Aviation Psychology Program Research Report No. 4., U.S. Government Printing Office, Washington, DC, 1947.

[2] Youngling, E. W., Levine, S. H., Mocharnuk, J. B., and Weston, L. M., *Feasibility Study to Predict Combat Effectiveness for Selected Military Roles: Fighter Pilot Effectiveness*. TR-MDC E1634, McDonnell-Douglas Astronautics Co., East St. Louis, MO, 1977.

[3] Imhoff, D. L. and Levine, J.M., Perceptual-motor and Cognitive Performance Task Battery for Pilot Selection AFHRL-TR-80-27, Air Force Human Resources Laboratory, San Antonio, TX, 1981.

[4] Damos, D.L. and Gibb, G.D., *Development of a Computer-Based Naval Aviation Selection Test Battery*, NAMRL 1319, Naval Aerospace Medical Research Laboratory, Pensacola, FL, August 1986.

[5] Damos, D. L., *Some Considerations in the Design of a Computerized Human Information Processing Battery*, NAMRL Monograph 35, Naval Aerospace Medical Research Laboratory, Pensacola, FL, December 1987.

[6] Bordelon, P. V. and Kantor, J. E., *Utilization of Psychomotor Screening for USAF Pilot Candidates: Independent and Integrated Selection Methodologies*, AFHRL-TR-86-4, Air Force Human Resources Laboratory, San Antonio, TX, 1986.

[7] Carretta, T. R.; "USAF Pilot Selection and Classification Systems." *Aviation, Space, and Environmental Medicine*, Vol.60, pp.46-49, 1989.

[8] Dolgin, D.L., Shull, R.N., and Gibb, G.D., "Risk Assessment and the Prediction of Student Pilot Performance." *In Proceedings of the 4th International Symposium on Aviation Psychology*, pp.480-485, 1989.

[9] Dolgin, D. L. and Gibb, G. D., "Personality Assessment in Aviator Selection: Past, Present, and Future." In *R. Jensen (Ed.), Aviation Psychology: The International Contribution.*, Gower Publishing Group, London, pp.285-319, 1989.

[10] Gibb, G. D. and Dolgin, D. L., "Predicting Military Flight Training Success by a Compensatory Tracking Task." *Journal of Military Psychology*, Vol.1, pp. 235-240, 1989.

[11] Griffin, G.R., *Development and Evaluation of an Automated Series of Single and Multiple Dichotic Listening and Psychomotor Tasks*, NAMRL-1336, Naval Aerospace Medical Research Laboratory, Pensacola, FL 1987.

[12] Morrison, T.R., *Complex Visual Information Processing: A Test for Predicting Navy Primary Flight Training Success*, NAMRL-1338, Naval Aerospace Medical Research Laboratory, Pensacola, FL 1988.

[13] Shull, R. N. and Dolgin, D. L., "Personality and Flight Training Performance." *Proceedings of the Human Factors Society 33rd Annual Meeting*, Denver, CO, 16-20 October 1989, pp. 891-895, 1989.

[14] Shull, R. N., Dolgin, D. L., and Gibb, G. D., *The Relationship between Flight Training Performance, a Risk Assessment Test and the Jenkins Activity Survey*, NAMRL-1339, Naval Aerospace Medical Research Laboratory, Pensacola, FL 1988.

[15] Gopher, D. and Kahneman, D., "Individual Differences in Attention and their Prediction of Flight Criterion." *Perceptual and Motor Skills*, Vol.33, pp.1335-1342, 1971.

[16] Damos, D.L., *Cross-adaptive Measurement of Residual Attention to Predict Pilot Performance.* University of Illinois at Urbana-Champaign, Aviation Research Laboratory Report No. TR ARL-72-25/AFOSR-72-14, Savoy IL, 1973.

[17] North, R.A. and Gopher, D., "Measures of Attention as Predictors of Flight Performance." *Human Factors*, Vol. 18, pp.1-14, 1976.

[18] Berg, C., Hertzog, C., and Hunt, E., "Age Differences in the Speed of Mental Rotation." *Developmental Psychology*, Vol. 18, pp.95-107, 1982.

[19] Carter, R. and Wolstad, J., "Repeated Measurements of Spatial Ability with the Manikin Test." *Human Factors*, Vol. 27, pp.209-220, 1985.

[20] Baddeley, A.D., "A Three-minute Reasoning Test based on Grammatical Transformation." *Psychonomic Science*, Vol. 10, pp.341-342, 1968.

[21] Slovic, P., "Manipulating the Attractiveness of a Gamble without Changing its Expected Value." *Journal of Experimental Psychology*, Vol. 79, pp.139-145, 1969.

[22] Cohen, J. and Cohen, P. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Lawrence Erlbaum, Hillsdale, NJ 1975.

[23] Tabachnick, B. and Fidell, L. *Using Multivariate Statistics.*, 2nd ed., Harper & Row, New York, 1989.

[24] Montgomery, D. and Peck, E. *Introduction to Linear Regression Analysis.*, John Wiley, New York 1982.

[25] Cox, D. R. *The Analysis of Binary Data.*, Methuen, London 1970.

[26] Carretta, T. R., *Cross-validation of Experimental USAF Pilot Training Performance models.*, AFHRL-TR-89-68, Air Force Human Resources Laboratory, San Antonio, TX 1990.

[27] Byrnes, P. *Estimates of Success Rates in the Aviator Training Pipeline.*, Research Memorandum, CRM89–73, Center for Naval Analyses, Washington, DC, August 1989.

APPENDIX A

The constituent tests of the performance-based test battery, their duration, and a brief description of the main cognitive attributes being measured, are shown below.

Test Name	TestLength (in min)	Attribute(s) Measured
Test Battery Introduction	15	biographical
Psychomotor Test (PMT)	41	eye-hand-foot coordination
Dichotic Listening Task (DLT)	15	divided attention
Combined PMT/DLT	18	time-sharing; multi-task performance
Manikin	10	visual/spatial; RT
Horizontal Tracking (HT)	20	compensatory tracking skills
Absolute Difference (AD)	30	memory; RT
Combined AD and HT	25	time-sharing; tracking ability
Complex Visual Task (CVT)	50	information processing from visual symbols
Baddeley Grammatical Reasoning	10	logical reasoning
Risk test	10	risk taking
TOTAL	244	

APPENDIX B

This appendix contains a description of the raw data base examined in this report. It also describes how the predictor variables used in the hierarchical multiple regression were derived.

The data base for the current report consisted of student naval aviators tested at the Naval Aerospace Medical Research Laboratory (NAMRL) on an Apple IIe system on one or more tests in the current battery. To be included in the current data base subjects had to have criterion data in the form of a pass/fail score from primary flight training. The 1110 subjects meeting this criterion were tested at NAMRL from April 1986 through March 1989.

Initially, the data base included information on 144 variables. (There was, naturally, a substantial amount of missing data as individual tests in the battery were introduced at different points over the period of testing.) The variables analyzed in the current report result from summarizing this information: combining some variables, transforming others, and ignoring yet others.

The 144 variables could be partitioned into 3 categories: 20 background and AQT/FAR variables, 116 measures from the computerized battery, and 8 outcome variables. Each of these categories will be described separately.

BACKGROUND AND AQT/FAR VARIABLES

As noted in the body of the current report, five background variables were chosen for examination: age, sex, accession source, college major, and previous flight hours. The latter three demographic variables were transformed so that they could be properly used in the regression equations. Dummy variable coding was employed for the categorical variables of accession source and college major.

Accession source was initially coded as a six-level variable with the following categories:

1. AOC/OCS
2. US Naval Academy
3. NAVCAD
4. Marine Corps
5. ROTC
6. Other

These initial six categories were reduced to three.

1. US Naval Academy
2. AOC/OCS
3. ROTC, NAVCAD, Marine Corps, and others

Two dummy variables, d_1 and d_2, were used for the variable accession source with the coding shown below.

Accession Source	d_1	d_2
Naval Academy	0	0
AOC/OCS	1	0
Others	0	1

Education, initially a five-level coding of college major, was also collapsed into three categories for the regressions. The dummy variable coding for college major is shown below.

College Major	d_1	d_2
General Science	0	0
Engineering/Math	1	0
Others	0	1

Previous flight hours had a mean of 24.24 but ranged from 0 to 2500 in the current data base. Because of the extreme skewness (12.04) of this variable, a logarithmic transformation to the base 10 of previous flight hours was used in the regression analyses.

Six AQT/FAR variables were selected for inclusion. Stanine scores for Academic Qualification Test (AQT), Flight Aptitude Rating (FAR), Spatial Apperception Test (SAT), Mechanical Comprehension Test (MCT), Biographical Inventory (BI), and the Officer Aptitude Rating (OAR) were used.

OUTCOME MEASURES

For this report, only one outcome measure was selected from the eight available. Failing or passing in primary training was coded as 0 and 1, respectively. Failures at intermediate and advanced stages of training were treated as a pass given that our focus was on primary training.

MEASURES FROM THE COMPUTERIZED BATTERY

The 116 variables from the computerized battery were reduced first to 38 and then to 25 summary measures. In addition to the six tasks described in the body of the report a pilot personality inventory (PPI) was also included in the battery relatively late in the testing period. The initial number of variables for the various tasks was as follows:

Task	Initial Number of Variables
CVT	3
Risk	2
ADHT	54
Manikin	24
Baddeley	6
PMT/DLT	15
PPI	12
Total	116

The variables for the CVT, Risk, and PPI were analyzed without any changes. The ADHT, Manikin, and to some extent the PMT/DLT had more measures because the same measures were recorded on each of several replications of the same task. Further, the absolute difference portion of the ADHT, as well as the Manikin and Baddeley tasks, included sets of six measures for a given replication: number of correct responses, mean and standard deviation of reaction times for correct responses, number of errors, and mean and standard deviation of reaction times for erroneous responses.

The general strategy was to make some *a priori* decisions about promising variables to achieve a set of predictors of more manageable size for analysis. First, for more reliable predictors, performance was averaged when the same measures were recorded for multiple replications of the same task. Second, because errors were fewer than correct responses, the mean reaction times for errors was less stable than the mean reaction times for correct responses. Consequently, error reaction times were ignored as were the standard deviations of reaction times. Third, as a timed test, the Manikin test produced reaction times that were largely redundant with number of responses so they were excluded from the final analyses. Finally, since the original data from the PMT task included partially redundant information on the most complex task (*i.e.*, the stick, rudder and throttle task), a single composite score of the nonredundant portions of performance was used. These reductions yielded 38 summary measures, including 8 for ADHT, 8 for PMT/DLT, 2 for the Manikin, and 3 for the Baddeley, together with the original 2 for Risk, 3 for CVT, and 12 for PPI.

Preliminary analyses suggested some further modifications and reductions. First, multiple regressions of both flight grade and pass/fail on the 12 PPI subscales proved to be nonsignificant. Further, data were available on these variables for only 78 subjects, whereas all other tasks had at least 337 subjects. Thus, the personality inventory was not included in the major analysis. Second, previous analyses of the PMT/DLT aviator

data and preliminary analyses of the current data indicated that the PMT tracking errors were very positively skewed, and the DLT correct scores were very negatively skewed. Thus, logarithmic transformations of number of errors were used for both variables to achieve more nearly normal distributions and more nearly linear relationships with the criteria. Finally, the most complex PMT task was dropped because it was available for only 250 subjects whereas the other portions were available for over 700 subjects. These transformations yielded the final set of 25 summary measures for the performance-based tests as summarized below:

Task	Final Number of Variables
CVT	3
Risk	2
ADHT	8
Manikin	2
Baddeley	3
PMT/DLT	7
PPI	0
Total	25

An enumeration of the final set of variables is contained in the following two tables.

Background Variables	
AGE	age in years at time of testing
SEX	gender
ACCESS	accession source
EDUC	college major
PFLTH	$\log_{10}$ of previous flight hours

AQT/FAR Variables	
AQT	Academic Qualifications Test stanine
FAR	Flight Aptitude Rating stanine
SAT	Spatial Apperception Test stanine
MCT	Mechanical Comprehension Test stanine
BI	Biographical Inventory stanine
OAR	Officer Aptitude Rating
Outcome Measures	
PF	Pass/Fail for primary training
Summary Measures from Computerized Tests	
CVT1	CVT: number correct responses
CVT2	CVT: mean time taken to read question
CVT3	CVT: mean time taken to enter answer
RISK1	Risk: number of boxes chosen in first 10 trials
RISK2	Risk: mean reaction time to choose box, 10 trials
ADHT1	ADHT: mean tracking errors trials 1-3, single mode
ADHT2	ADHT: mean number correct, absolute differences, trials 1-5, single mode
ADHT3	ADHT: mean number errors, absolute differences, trials 1-5, single mode
ADHT4	ADHT: mean reaction time, correct responses, absolute differences trials 1-5, single mode
ADHT5	ADHT: mean tracking errors trials 1-3, dual mode
ADHT6	ADHT: mean number correct, absolute differences, trials 1-3, dual mode
ADHT7	ADHT: mean number errors, absolute differences, trials 1-3, dual mode
ADHT8	ADHT: mean reaction time, correct responses, absolute differences, trials 1-3, dual mode
MAN1	Manikin: mean number correct, trials 1-4
MAN2	Manikin: mean number errors, trials 1-4
BAD1	Baddeley: number correct
BAD2	Baddeley: mean reaction time, correct responses
BAD3	Baddeley: number errors
PMTDLT1	PMT/DLT: $\log_{10}$ (dichotic listening errors + 1), single mode
PMTDLT2	PMT/DLT: $\log_{10}$ (dichotic listening errors + 1), dual mode (with PMT stick)
PMTDLT3	PMT/DLT: $\log_{10}$ (dichotic listening errors + 1), dual mode (with PMT stick & rudder)
PMTDLT4	PMT/DLT: $\log_{10}$ tracking errors, stick, single mode
PMTDLT5	PMT/DLT: $\log_{10}$ tracking errors, stick, dual mode
PMTDLT6	PMT/DLT: $\log_{10}$ tracking errors, stick & rudder, single mode
PMTDLT7	PMT/DLT: $\log_{10}$ tracking errors, stick & rudder, dual mode

Descriptive statistics for these 25 measures and for the 6 AQT/FAR variables complete this appendix. The number of subjects (N) may differ here from that given in the main text due to deletion of cases not containing values on all variables for a given regression. For example, $N = 1110$ for the AQT/FAR variables in the table below. But there were only 1077 cases where data existed on all 6 AQT/FAR variables simultaneously.

AQT/FAR Variables

Variable	Mean	*SD*	Minimum	Maximum	*N*
AQT	5.60	1.26	2	9	1110
FAR	7.06	1.58	2	9	1110
SAT	12.74	3.10	1	19	1110
MCT	11.50	2.86	2	19	1110
BI	13.25	3.28	2	19	1110
OAR	50.30	7.27	21	80	1110

CVT and Risk Variables

Variable	Mean	*SD*	Minimum	Maximum	*N*
CVT1	100.14	9.43	8.00	117.00	562
CVT2	8.23	1.93	4.21	16.72	562
CVT3	6.01	1.42	3.21	11.07	562
RISK1	4.74	.91	1.70	7.30	340
RISK2	3.82	1.93	1.23	14.03	340

Absolute Difference-Horizontal Tracking Variables

Variable	Mean	*SD*	Minimum	Maximum	*N*
ADHT1	30.12	12.54	3.40	85.51	529
ADHT2	73.35	16.60	32.60	127.40	525
ADHT3	11.35	24.59	0.00	301.00	525
ADHT4	1.91	.37	.41	4.10	525
ADHT5	35.79	14.99	4.21	101.62	514
ADHT6	65.77	17.35	22.67	127.33	514
ADHT7	10.79	21.57	.33	377.67	514
ADHT8	1.99	.43	.25	4.99	514

Manikin and Baddeley Variables

Variable	Mean	*SD*	Minimum	Maximum	*N*
MAN1	25.42	5.24	9.50	39.00	563
MAN2	1.43	2.07	0.00	20.50	563
BAD1	58.51	15.67	13.00	114.00	587
BAD2	3.91	1.79	1.09	20.05	587
BAD3	6.67	7.54	0.00	85.00	587

Psychomotor/Dichotic Listening Test Variables

Variable	Mean	*SD*	Minimum	Maximum	*N*
PMTDLT1	.75	.27	0.00	1.79	699
PMTDLT2	.84	.36	0.00	1.86	699
PMTDLT3	.92	.30	0.00	1.93	701
PMTDLT4	4.05	.28	3.52	5.20	711
PMTDLT5	3.62	.28	3.13	4.81	708
PMTDLT6	4.56	.22	4.10	5.53	705
PMTDLT7	4.00	.25	3.39	4.95	685